Action recognition based on shape features and their correlation

Katarzyna Gościewska
Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Szczecin, Poland
kgosciewska@wi.zut.edu.pl

Abstract: This paper presents an approach for action recognition based on binary silhouette sequences extracted from consecutive frames of a video. It uses shape descriptors and correlation coefficient to represent and match entire sequences, regardless the number of frames. Each set of binary silhouettes corresponds to one action, such as jumping or waving. The paper provides experimental results on the use of the proposed approach and four shape description algorithms, namely the Two-Dimensional Fourier Descriptor, Generic Fourier Descriptor, Point Distance Histogram and UNL-Fourier Descriptor. The results are analysed in terms of the highest classification accuracy and the smallest shape descriptor size.

Keywords: action recognition, shape descriptors, correlation coefficient, silhouette sequences

1. Introduction

A single action is a type of activity composed of multiple temporarily organized gestures performed by a single person, while a gesture is an elementary body movement executed for a short time [1]. Running, walking or bending are exemplary actions. According to [2], every human movement can be represented as a continuous pose change, and silhouettes extracted from consecutive video frames can be used for action classification based on shape features.

In the literature one can find a number of shape-based methods that employ representations of silhouette sequences for action recognition. For instance, the authors of [3] take an individual silhouette and convert it into a one-dimensional representation which is further transformed into symbolic vector called Symbolic Aggregate approXimation (SAX). Then a set of SAX vectors is used as the representation of one action. In turn, the authors of [4] proposed a History Trace Template for action representation which is composed of the set of Trace Transforms extracted for each silhouette. There are also other approaches which limit the number of silhouettes to characteristic frames (key poses), e.g. [5,6] or [7] where Dynamic Time Warping is used to match sequences of key poses, or accumulates silhouettes in the space-time volume where actions are represented as three dimensional shapes, e.g. [8].

This paper concerns an action recognition approach that uses shape descriptors and correlation coefficient to classify actions based on sets of consecutive binary silhouettes extracted from video sequences. For a single video sequence, each extracted silhouette is represented using a particular shape description algorithm, and the representation of the first frame is matched with the rest of representations using correlation coefficient. Obtained
correlations are put into one vector, which is further processed and transformed into a
sequence representation. Then, template matching approach, with the variable template set,
is iteratively used to match sequence representations and investigate the average
classification accuracy. Detailed data processing steps of the developed approach are
explained in Section 2. Section 3 describes selected shape description algorithms, Section 4
presents some experimental results and Section 5 concludes the paper.

2. A description of the developed approach

The proposed approach enables to create representations of entire sequences where one
sequence is composed of a set of binary images containing silhouettes of foreground objects
extracted from the consecutive frames of the particular video sequence showing a person
performing an action. The following description explains how individual silhouettes are
represented and how they are processed to obtain a final sequence representation. The
matching procedure and the manner of result analysis are also given.

2.1. Shape representation and matching

Each silhouette is represented using one shape description algorithm. Contour or region-
based methods can be used. Some algorithms enable to calculate shape descriptors of differ-
ent size, for instance in case of the shape description algorithms based on the Fourier Tran-
sform we can use various subparts of the coefficient matrix. The size of shape representation
may influence final classification accuracy and processing time. Therefore we are trying to
select the smallest descriptor which carries the most information enabling shape differentia-
tion.

In the next step, correlation coefficient is used to calculate the similarity of the first
frame and the rest of frames based on shape descriptors. The resulting correlation values are
put into a vector. The vector is normalized to interval [0,1] and the number of its elements is
equal to the number of frames (silhouettes) in a particular sequence. At this moment a data-
base contains vectors, so called similarity signals, one vector for one sequence.

2.2. Sequence representation and matching

Due to the fact that similarity signals are of different length, they are subjected to further
processing. Firstly, a fast Fourier transform (FFT) algorithm is applied to a similarity signal
and only magnitude is taken. Then a periodogram is calculated – it enables to equalize
lengths of vectors. The combination of the FFT magnitude and periodogram gave the most
satisfactory results so far. Ultimately, a periodogram constitutes a sequence representation.

The final step of the proposed approach is to calculate classification accuracy using
template matching technique and correlation coefficient. Some initial tests demonstrated
that final results vary depending on the selected templates (one template represents one ac-
tion class). Therefore, each experiment is performed iteratively using different set of tem-
plates in each iteration. Then the final classification accuracy is calculated as an average of
all iterations – similarly as in the k-fold cross-validation technique. For instance, using the
database with n sequence representations and divided into k folds, the first iteration uses
sequence representations with numbers from 1 to k as templates and sequence representa-
tions of numbers from k+1 to n as test objects, then the second iteration uses sequence rep-
resentations with number from k+1 to 2*k as templates and the rest as test objects, and so
on. Sequence representations are matched using correlation coefficient and template match-
ing approach – a particular test object is matched with all templates and the most similar
template indicates to which class a processed test object belongs. The experimental classification results are compared to actual class labels and the percentage of correct classifications is obtained. Then the results can be analysed in three different ways:

- Overall classification accuracy calculated for each shape descriptor, averaged for all classes and averaged for all iterations.
- Classification accuracy calculated for each iteration and each shape descriptor, and averaged for all classes.
- Classification accuracy calculated for each class and each shape descriptor (using one iteration only or an average for all iterations).

3. Four selected shape description algorithms

The proposed approach has been experimentally investigated using two region-based shape description algorithms, namely the Two-Dimensional Fourier Descriptor and Generic Fourier Descriptor, as well as two methods based on contour points, i.e. Point Distance Histogram and UNL-Fourier Descriptor. Their definitions are given below.

The Two-Dimensional Fourier Descriptor is simple to obtain and robust to noise. The representation has a form of a matrix with absolute complex values and is calculated for the region shape. It is often sufficient to use only a small subpart of the original representation, and therefore in the experiments various subparts are be used. The original representation is derived using the formula [9]:

\[
C(k,l) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} P(h,w) \cdot e^{-\frac{2\pi i (h-1)(k-1)}{H}} \cdot e^{-\frac{2\pi i (l-1)(w-1)}{W}}
\]

where:

- \( H, W \) – height and width of the image in pixels,
- \( k \) – sampling rate in vertical direction \((k \geq 1 \text{ and } k \leq H)\),
- \( l \) – sampling rate in horizontal direction \((l \geq 1 \text{ and } l \leq W)\),
- \( C(k,l) \) – value of the coefficient of discrete Fourier transform in the coefficient matrix in \( k \) row and \( l \) column,
- \( P(h,w) \) – value in the image plane with coordinates \( h, w \).

The Generic Fourier Descriptor is also a region-based Fourier Descriptor that utilizes pixel transformation to polar coordinate system [10,11]. All pixel coordinates of an original region shape image are transformed into polar coordinates and new values are put into a rectangular Cartesian image. Elements in rows correspond to distances from centroid and the columns are corresponding to angles. In result, an image of a transformed shape is obtained and the Two-Dimensional Fourier Transform can be applied.

The first method based on contour points and polar transformation is Point Distance Histogram. In order to obtain basic representation, the centroid is calculated. Then the shape contour is transformed into polar coordinates, and new coordinates are put into two vectors – \( \Theta \) for angles and \( \rho \) for radii (the description below is based on [12,13]):

\[
\theta_i = a \tan \left( \frac{y_i - O_y}{x_i - O_x} \right) \quad \rho_i = \sqrt{(x_i - O_x)^2 + (y_i - O_y)^2}.
\]

Values in \( \Theta \) are converted into nearest integers. In the next step the elements of \( \Theta \) and \( \rho \) are sorted according to increasing values in \( \Theta \) and denoted as \( \Theta^I \) and \( \rho^I \). If there are some equal angle values in \( \Theta^I \), then only the value with the highest corresponding radii value in \( \rho^I \)
is left. These transformations result in a vector with no more than 360 elements, and only \( P^j \) is further processed (denoted as \( P^k \)). Vector \( P^k \) is normalized according to its highest value:

\[
M = \max_k \{\rho_k\}, \quad \rho_k = \frac{\rho_k}{M}.
\]  

(3)

The elements in \( P^k \) are assigned to bins in histogram (\( \rho_k \) to \( l_k \)):

\[
l_k = \left\lfloor r, \quad \text{gdy } \rho_k = 1 \right\rfloor [r \rho_k], \quad \text{gdy } \rho_k \neq 1
\]

(4)

where \( r \) is a previously defined number of histogram bins. In the next step, the values in bins are normalized according to the highest one:

\[
S = \max_k \{l_k\}, \quad l_k = \frac{l_k}{S}.
\]

(5)

The final histogram can be written as a function \( h(l_k) \):

\[
h(l_k) = \sum_{i=1}^{m} b(k,l_k),
\]

(6)

where:

\[
b(k,l_k) = \begin{cases} 1, & \text{gdy } k = l_k \\ 0, & \text{gdy } k \neq l_k \end{cases}
\]

(7)

The UNL-Fourier Descriptor is composed of UNL descriptor and the Two-Dimensional Fourier Transform. The UNL utilizes complex representation of Cartesian coordinates for points and parametric curves in discrete manner [14]:

\[
z(t) = (x_1 + t(x_2 - x_1)) + j(y_1 + t(y_2 - y_1)), \quad t \in (0,1),
\]

(8)

where \( z_1 = x_1 + jy_1 \) and \( z_2 = x_2 + jy_2 \) are complex numbers. In the next step, the centroid \( O \) is calculated [14]:

\[
O = (O_x, O_y) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right).
\]

(9)

where \( n \) is the number of points in a contour, and \( x_i, y_i \) are Cartesian coordinates of the \( i \)-th point. The maximal Euclidean distance between contour points and centroid is found [14]:

\[
M = \max_i \|z_i(t) - O\| \quad \forall i = 1...n \quad t \in (0,1),
\]

(10)

New coordinates in a discrete version can be derived using the following formula [14]:

\[
U(z(t)) = \left\lfloor \frac{1}{M} \left[ (x_1 + t(x_2 - x_1) - O_x) + j(y_1 + t(y_2 - y_1) - O_y) \right] \right\rfloor + j \times \text{atan} \left( \frac{y_1 + t(y_2 - y_1) - O_y}{x_1 + t(x_2 - x_1) - O_x} \right).
\]

(11)

The parameter \( t \) is discretized in the interval \([0,1]\). Values of new coordinates are put into a matrix – rows represent distances from the centroid and columns the corresponding angles. This results in a Cartesian image containing unfolded shape contour as it is seen in polar coordinates. Then the Two-Dimensional Fourier Transform is applied, what makes the representation invariant to rotation.
4. Experimental conditions and results

4.1 Data and conditions

The proposed approach has been experimentally evaluated using a half of the Weizmann database [15]. The entire database consists of 90 short video sequences (180 x 144, 50 fps) depicting 9 actors performing 10 different actions. The number of frames (silhouettes) in a sequence varied from 28 to 125. The selected half of the database contains silhouette sequences of bending, jumping, running, walking and one-hand waving actions (see Figure 1 for exemplary silhouettes). The reason for choosing these actions was their popularity in real human behaviours compared to the rest of actions in the database as well as the higher possibility that the selected action would occur during visual content analysis in video surveillance systems or would be an element of more complex behaviour (overall usefulness).

Several experiments have been carried out to verify the proposed approach using different shape description algorithms, namely the Two-Dimensional Fourier Descriptor, Generic Fourier Descriptor, Point Distance Histogram and UNL-Fourier Descriptor. One experiment included several tests using various size of shape representations obtained using a particular shape description algorithm. The 2x2, 5x5, 10x10, 25x25 and 50x50 absolute spectrum subparts for the Fourier-based methods and 2, 5, 10, 25 and 50 histogram bins for the Point Distance Histogram were taken. As the number of sequences in the database equalled to 45, then each subgroup of 5 sequences of different actions performed by one actor was iteratively used as a template set. Subsection 3.2 presents percentage experimental results.

Figure 1. Exemplary silhouettes extracted from video sequences Weizmann dataset [15] – silhouettes in rows correspond to (from the top) bending, jumping, running, walking and waving actions.
4.2 Experimental results

Overall classification accuracy calculated for each shape descriptor, averaged for all classes and averaged for all iterations are as follows:

- 52.5%, 51.7%, 50.3%, 51.6% and 49.7% for 2x2, 5x5, 10x10, 25x25 and 50x50 absolute spectrum subparts of the Two-Dimensional Fourier Descriptor respectively;
- 44.4%, 50.8%, 52.2%, 53.1% and 52.8% for the 2x2, 5x5, 10x10, 25x25 and 50x50 absolute spectrum subparts of the Generic Fourier Descriptor respectively;
- 33.9%, 49.7%, 51.1%, 50.3% and 48.1% for the 2x2, 5x5, 10x10, 25x25 and 50x50 absolute spectrum subparts of the UNL-Fourier Descriptor respectively;
- 31.1%, 51.1%, 50.8%, 48.9% and 43.9% for the 2, 5, 10, 25 and 50 histogram bins of the Point Distance Histogram respectively.

In general, the averaged classification accuracy results indicated the Two-Dimensional Fourier Descriptor as the most effective shape representation method to be employed in the proposed approach. In this case accuracy values for different descriptor sizes were very similar in contrast to the other experiments for which the use of the smallest shape descriptors gave noticeably poorer results. This may be due to different characteristics of the methods – the Two-Dimensional Fourier Descriptor is the only one which is not rotation-invariant (initial shape is not transformed into polar coordinates). Comparing each test individually, the highest accuracy values were obtained with the use of the 25x25 and 50x50 subparts of the Generic Fourier Descriptor – accuracy equal to 53.1% and 52.8% respectively – and then by the 2x2 and 5x5 subparts of the Two-Dimensional Fourier Descriptor – accuracy equal to 52.5% and 51.7% respectively. However, given that these are only averaged values, we have to take into account accuracies of individual iterations as well as the size of each descriptor. According to that, the best solution is the use of the 2x2 subpart of the Two-Dimensional Fourier Descriptor – it uses the smallest representation size and additionally accuracy values obtained in iteration no. 6 exceeds averaged results as well as accuracies obtained in individual iterations in the experiment applying the Generic Fourier Descriptor.

Table 1 presents classification accuracy values calculated for each iteration and each shape descriptor, and averaged for all classes in the experiment using the Two-Dimensional Fourier Descriptor. In turn, Table 2 contains classification accuracy calculated for each class and each shape representation size of the same experiment. It can be seen that classification accuracy varies depending on the class – ‘bend’ action is most recognizable, while ‘jump’ and ‘walk’ actions are less distinctive ones.

Table 1. Results for the experiment using the Two-Dimensional Fourier Descriptor—classification accuracy for each iteration, each shape descriptor and averaged for all classes.

<table>
<thead>
<tr>
<th>Iteration no.</th>
<th>Size of the Two-Dimensional Fourier Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>1</td>
<td>42.5%</td>
</tr>
<tr>
<td>2</td>
<td>55.0%</td>
</tr>
<tr>
<td>3</td>
<td>50.0%</td>
</tr>
<tr>
<td>4</td>
<td>50.0%</td>
</tr>
<tr>
<td>5</td>
<td>42.5%</td>
</tr>
<tr>
<td>6</td>
<td>67.5%</td>
</tr>
<tr>
<td>7</td>
<td>45.0%</td>
</tr>
<tr>
<td>8</td>
<td>62.5%</td>
</tr>
<tr>
<td>9</td>
<td>57.5%</td>
</tr>
</tbody>
</table>
Table 2. Results for the experiment using the Two-Dimensional Fourier Descriptor—classification accuracy for each class, each shape descriptor and averaged for all iterations.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size of the Two-Dimensional Fourier Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>'bend'</td>
<td>95,8%</td>
</tr>
<tr>
<td>'jump'</td>
<td>31,9%</td>
</tr>
<tr>
<td>'run'</td>
<td>62,5%</td>
</tr>
<tr>
<td>'walk'</td>
<td>31,9%</td>
</tr>
<tr>
<td>'wave'</td>
<td>40,3%</td>
</tr>
</tbody>
</table>

Let us focus only on iteration no. 6 of the experiment using 2x2 subpart of the Two-Dimensional Fourier Descriptor. Figure 2 depicts plots of similarity signals (normalized vectors with between-frame correlation coefficient values) corresponding to five templates (sequence representations) used as class representatives in iteration no. 6. The plots show differences between actions and indicate which actions are periodical. High peaks on the plots correspond to the silhouettes that are most similar to the first silhouette in a particular sequence. Table 3 contains classification accuracy values obtained in the experiment using the 2x2 subpart of the Two-Dimensional Fourier Descriptor and iteration no. 6. The results are mostly better than averaged, probably due to the most distinctive templates.

![Similarity signals](image)

Figure 2. Five similarity signals corresponding to sequence representations used as templates during iteration no. 6 of the experiment using 2x2 subpart of the Two-Dimensional Fourier Descriptor.

Table 3. Results for the experiment using the Two-Dimensional Fourier Descriptor—classification accuracy for each class and iteration no. 6.

<table>
<thead>
<tr>
<th>Class</th>
<th>Size of the Two-Dimensional Fourier Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>'bend'</td>
<td>100%</td>
</tr>
<tr>
<td>'jump'</td>
<td>50%</td>
</tr>
<tr>
<td>'run'</td>
<td>50%</td>
</tr>
<tr>
<td>'walk'</td>
<td>75%</td>
</tr>
<tr>
<td>'wave'</td>
<td>62,5%</td>
</tr>
</tbody>
</table>
4. Conclusions

In the paper, an action recognition approach based on shape descriptors and correlation coefficient has been presented. It uses binary silhouettes extracted from the consecutive frames of a video sequence as input data. Each silhouette is represented using shape description algorithm and the representation of the first frame is matched with the rest of representations using correlation coefficient. Obtained correlation values are put into a vector which is further processed to prepare a final sequence representation—normalization, fast Fourier transform and periodogram are applied. Action classification employs template matching technique and correlation coefficient. The most satisfactory results in terms of the highest classification accuracy and the smallest descriptor size has been obtained using the 2x2 subpart of the Two-Dimensional Fourier Descriptor and a template set applied in iteration no. 6.

References